PATENT ABSTRACTS OF JAPAN

(11)Publication number : 08-241332
(43)Date of publication of application : 17.09.1996

(51)Int.CI. G06F 17/30

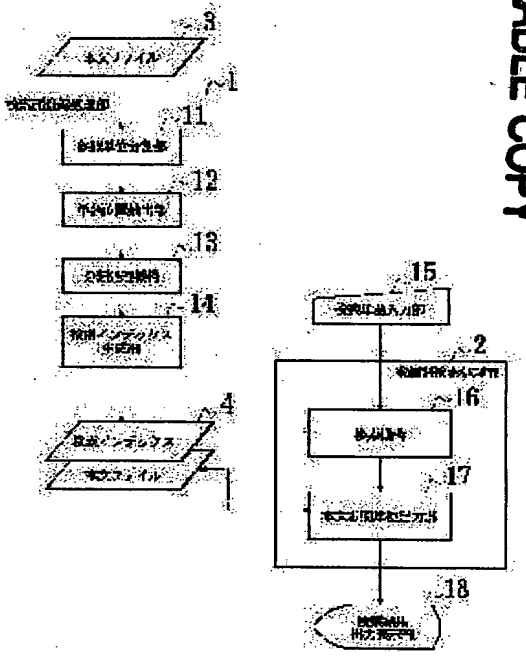
(21)Application number : 07-066727
(22)Date of filing : 02.03.1995

(71)Applicant : FUJI XEROX CO LTD
(72)Inventor : TATENO SHOICHI

(54) DEVICE AND METHOD FOR RETRIEVING ALL-SENTENCE REGISTERED WORD

(57)Abstract:

PURPOSE: To provide the device and method for retrieving all-sentence registered word which can obtain reference words of the main body of a tagged document by regarding tags as reference units for retrieval results and efficiently retrieving word positions in the main body.



CONSTITUTION: The device and method are equipped with a reference unit division part 11 which inputs a main body file 3 containing the main body of the tagged document and divides it into the reference units sectioned with the tags, a word position extraction part 12 which extracts pairs of words, included in the reference units, to be retrieved and the positions of the reference units in the main body where the words appear, a classification pat 13 which classifies the pairs of the extracted words and the positions of the reference units according to the words and obtains a word position set of the positions of all the reference units where the words appear paired with the words, and a retrieval index generation part 14 which generates retrieval indexes for obtaining the position set from the words for the word position set.

LEGAL STATUS

[Date of request for examination]	18.10.1996
[Date of sending the examiner's decision of rejection]	
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	2896634
[Date of registration]	12.03.1999
[Number of appeal against examiner's decision of rejection]	
[Date of requesting appeal against examiner's	

1

【特許請求の範囲】

【請求項１】　タグを有する文書の本文を収めた本文ファイルを入力し、タグで区切られた参照単位に分割する参照単位分割部と、
参照単位に含まれる検索対象とする単語に対して、単語と当該単語が出現する本文における参照単位の位置の対を抽出する単語位置抽出部と、
抽出された単語と参照単位の位置の対を単語に従って分類し、単語に対し当該単語が出現する全ての参照単位の位置を組とした単語位置集合を得る分類部と、
単語位置集合に対し、単語から位置集合を得る検索インデックスを生成する検索インデックス生成部とを備えることを特徴とする全文登録語検索装置。

【請求項２】　タグを有する文書の本文を収めた本文ファイルをタグで区切られた参照単位に分割し、
参照単位内に含まれる検索対象とする単語に対して、単語と当該単語が出現する全ての参照単位の位置の対を抽出し、
抽出され単語と参照単位の位置の対を単語により分類し、
単語と当該単語が出現する全ての参照単位の位置を組とした単語位置集合を作成し、
作成された単語位置集合に基づいて、単語から位置集合を得ることができる検索インデックスを生成することを特徴とする全文登録語検索方法。

【請求項３】　請求項１に記載の全文登録語検索装置において、更に、
検索インデックス生成部により作成された検索インデックスを用いて得られたタグの位置の集合に基づいて、参照単位を次のタグまであるいは適当な長さだけ表示して、検索結果を表示する検索処理部を有することを特徴とする全文登録語検索装置。

【請求項４】　請求項３に記載の全文登録語検索装置において、
タグに付加されたフィールド情報に基づいて特定のフィールドを検索対象とする場合、
本文中の単語について、その単語の直前にフィールドの種別を示す文字列を付加したものを新たな単語とし、当該単語が出現する位置の直前にあるタグの位置との対を単語位置集合対とし、
単語を入力し、その語が出現する位置の直前にあるタグの位置の集合を結果として出力することを特徴とする全登録語検索装置。

【請求項５】　請求項３に記載の全文登録語検索装置において、
本文中に指定形式で記述された属性と値の対を含む場合、
前記属性と値の対を単語として登録し、
検索時に指定形式の一部を入力することにより、検索対象の文字列から始まる指定形式の属性と値の対を列挙することを特徴とする全文登録語検索装置。

2

【請求項６】　請求項３に記載の全文登録語検索装置において、
特定のフィールドを検索対象とし、本文中に指定形式で記述された属性と値の対を含む場合、
属性と値の対の直前にフィールドの種別を示す文字列を付加したものを単語として登録し、
検索時に指定形式の一部を入力することにより、検索対象の指定形式の文字列から始まる指定形式を列挙することを特徴とする全文登録語検索装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、全文登録語検索装置および方法に関し、特に、タグを有する文書において本文中の単語を登録し、全文の登録語の検索を能率よく行う全文登録語検索装置および方法に関するものである。

【０００２】

【従来の技術】従来から、ワークステーション上の文書編集装置（ワードプロセッサ）など、文書処理を行う文書編集装置においては、文書の作成を効率よく行うため、タグを用いて文書内容を部分的に区別して、予じめ、見出し，段落などの複数の文書部品を作成し、その各々の文書部品の間の関係を定めて、文書を構造化して編集することが試みられている。

【０００３】このような文書に対して構造の概念を取り入れた構造化文書の例としては、例えば、国際規格のＯＤＡ（ＩＳＯ８６１３：Open Document Architecture）や、ＳＧＭＬ（ＩＳＯ８８７９：Standard Generalized Markup Language）の規格による構造化文書が知られている。ＯＤＡの規格による構造化文書を用いた文書処理方法の一例は、特開平５－１３５０５４号公報に記載されている「文書処理方法」が参照できる。

【０００４】ところで、ＳＧＭＬによる構造化文書は、従来のテキスト処理システムとの親和性が高く、アメリカを中心として普及してきており、既に実用の段階に入っている。このようなＳＧＭＬによる構造化文書の手法では、タグとよばれるマークを文書テキスト中に挿入することで、文書テキストを部分的に区分し（例えば文書部品として区分し）、その区分の関係を規定して構造化し、木構造の文書構造を表現する。

【０００５】ＳＧＭＬによる構造化文書を例にとって、マーク（タグ）付けされた構造化文書の処理例について説明する。ＳＧＭＬによる構造化文書では、予じめ文書の構造のひな型が与えられ、文書の構造は、その与えられたひな型の範囲内に制約される。この文書構造のひな型は、ＳＧＭＬにおいては、文書型定義（ＤＴＤ：Document Type Definition）と呼ばれる。

【０００６】ＳＧＭＬの構造化文書では、まず、文書型定義を規定して、文書の構造を表現するために、文書テキスト内にタグと呼ばれるマークを挿入し、そのタグに

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平8－241332

(43)公開日　平成８年(1996)９月17日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/30		9194－5L	Ｇ０６Ｆ	15/401	３１０Ａ
		9194－5L		15/40	３７０Ａ

審査請求　未請求　請求項の数６　ＦＤ　（全 17 頁）

(21)出願番号　特願平7－66727

(22)出願日　平成７年(1995)３月２日

(71)出願人　000005496
富士ゼロックス株式会社
東京都港区赤坂二丁目17番22号

(72)発明者　舘野　昌一
神奈川県横浜市保土ヶ谷区神戸町134番地
横浜ビジネスパークイーストタワー　富士ゼロックス株式会社内

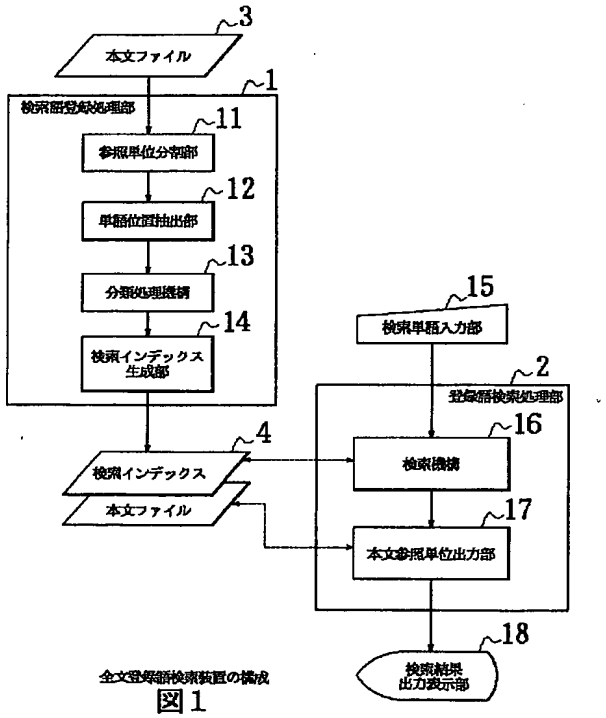
(74)代理人　弁理士　南野　貞男　（外２名）

(54)【発明の名称】　全文登録語検索装置および方法

(57)【要約】

【目的】　タグ付き文書において、タグを検索結果の参照単位として、効率よく本文中の単語位置を検索し、本文の参照単語を得ることができる全文登録語検索装置および方法を提供する。

【構成】　タグを有する文書の本文を収めた本文ファイルを入力し、タグで区切られた参照単位に分割する参照単位分割部と、参照単位に含まれる検索対象とする単語に対して、単語と当該単語が出現する本文における参照単位の位置の対を抽出する単語位置抽出部と、抽出された単語と参照単位の位置の対を単語に従って分類し、単語に対し当該単語が出現する全ての参照単位の位置を組とした単語位置集合を得る分類部と、単語位置集合に対し、単語から位置集合を得る検索インデックスを生成する検索インデックス生成部とを備える。



全文登録語検索装置の構成
図１

より文書テキストを部分的に区分する。例えば、文書における一つの段落は、名前が“段落”とされたタグ<段落>を用いて、次のように表現される。

『<段落>これは一つの段落です。</段落>』

ここでのタグ<段落>が段落の開始を意味し、スタートタグと呼ばれる。タグ</段落>が段落の終了を意味し、エンドタグと呼ばれる。つまり、ここでは、タグの名前が“段落”とされたスタートタグ<段落>とエンドタグ</段落>との２つのタグを用いてマーク付けし、文書におけるテキストを文書部品として部分的に区分する。つまり、２つのタグの間に挟まれたテキスト部分が、タグで指示された構造の内容部分を示している。

【０００７】名前が付けられたタグは各々が区別されて、文書型定義の中で構造上のその位置付けが定義される。その意味では、タグが文書の構造（構成要素）を表現している。したがって、混合が生じない場合において、以下で言う構造化文書（ＳＧＭＬによる文書）の構造とは、タグと同義であることを意味している。

【０００８】また、ＳＧＭＬによる構造化文書（以下、ＳＧＭＬ文書と略称する）においては、一部のタグを省略できる。その場合の省略の可／不可は、文書型定義（ＤＴＤ）により指定する。省略はスタートタグおよびエンドタグのそれぞれに独立に指定できる。例えば、エンドタグ</段落>が省略可とする場合は、これが文書型定義内で指定された場合であり、その場合、先の例は

『<段落>これは一つの段落です。』

と記述してもよいことになる。

【０００９】ＳＧＭＬ文書の文書型定義の具体例は、例えば、図１０に示される。図１０に示す文書型定義１００により規定される文書構造では、名前が“題”とされたスタートタグ、“題”のエンドタグ、“段落”のエンドダク、“図”のエンドタグ、および“図本体”のエンドタグが省略可能であることが定義されている。

【００１０】更に、図１０に示す文書型定義１００の内容を具体的に説明すると、ここでの文書型定義（ＤＴＤ）はＳＧＭＬの表記法に従って記述されているので、それに従って解釈できる。つまり、文書型定義の内容の行の最初の“<！”はマークアップ宣言区切り子であり、空白なしに続く次の“ＥＬＥＭＥＮＴ”は要素宣言キーワードである。この行の最初の“<！ＥＬＥＭＥＮＴ”により、次に続く記述によって、その構造の内容（下部の構造）がどのようになるのかを指定する。そして、その次に記述される項目の名前（文書，章，題，段落，図など）が、対象となるタグの名前を表している。

【００１１】更に、次の記号（“－　－”，“－　Ｏ”，“Ｏ　Ｏ”など）は、その項目の対象のタグが、スタートタグおよびエンドタグの順でそれぞれ省略可能かどうかを表す記号である。“－”が省略不可を意味し、“Ｏ”が省略可を意味する。例えば、ここでの記号が“－　Ｏ”であれば、スタートタグは省略不可であり、エンドタグは省略可であることを意味する。

【００１２】更に続く次の項目は、タグの下部の構造を表す規定の定義である。ここでの記号“，”は項目（タグ）が順序立てて出現することを意味し、記号“｜”はどちらかの項目であれば良いことを意味し、記号“＊”は０回以上の繰り返しを意味する。また、記号“？”はそれがあってもなくても良いことを意味している。

【００１３】したがって、例えば、タグの下部構造が“（章題，段落＊，章＊）”と規定されている場合は『章題の次に段落の０回以上の繰り返しがあり、更にその次に章の０回以上の繰り返しがある』という順序で下部の構造が規定されることを意味する。具体例で説明すると、図１０に示す文書型定義１００の第２行目のように、タグの下部構造が“（題，（段落｜図）＊，章＊）”と規定されている場合、『章の次に段落または図の０回以上の繰り返しがあり、その次に章の０回以上の繰り返しがある』順序で下部の構造が規定されることを意味する。

【００１４】また、第３行目および第４行目に記述されているタグの下部構造の“＃ＰＣＤＡＴＡ”はＳＧＭＬの予約語の１つであり、構造の規定で、その内容が文字データであること意味している。したがって、文書型定義１００の例では、章を構成する「題」と「段落」のタグの下部には文字データが来ることを意味している。

【００１５】つまり、図１０に示す文書型定義（ＤＴＤ）の意味するところによる文書構造のひな型では、当該文書が、「章」の繰り返しからなる“<文書>”というタグから始まる文書であり、その「章」は、「題」の次に「段落」または「図」の０回以上の繰り返しがあり、更にその次に「章」の０回以上の繰り返しがあるものから構成されている。そして、ここでの「題」および「段落」が、文字データから構成される。

【００１６】更に細部の規定として、文書構造の「図」の内容は、「題」とそれに続く「図本体」から構成されると定義され、「図本体」は、例えば外部のイメージファイルを参照するので、下部構造を持たない（“ＥＭＰＴＹ”）と定義される。また、ここでの構造のタグの省略可／省略不可の指定は、「文書」および「章」はタグの省略不可であり、「段落」，「図」，および「図本体」は、エンドタグのみが省略可であり、「題」は両方のタグが省略可であるということが定義される。

【００１７】このような文書型定義に従っている実際の文書（以下、対象文書と呼ぶ）の例としては、例えば、図１１に示されているＳＧＭＬ文書１１０がある。なお、この図１１に示す文書の中では、文書の構造の深さに応じてインデンテーションを変えて表記しているが、この表記は、ここでの構造化文書の文書例の説明上、見やすくするために行っているものであり、実際の文書ではインデンテーションされないことが多い。

【００１８】図１１を参照すると、この例の構造化文書

5

のＳＧＭＬ文書１１０を見ると理解されるように、「章」を構成するその下位の構造の「題」を表すタグは、スタートタグおよびエンドタグは共にこの文書中には現れていない。しかし、実体的には、第２行目のタグ“＜章＞”とその内容部分の“ＳＧＭＬとは”との間のスタートタグ“＜題＞”が省略されている形となっている。なお、このようなタグが省略されているかどうかは、前述の文書型定義１００を参照しなければ分からない。したがって、対象文書は常にそれに従っている文書型定義とのセットでないと正確な構造を読み取ることができない。

【００１９】ＳＧＭＬ文書では、このようにタグの省略が行われるため、ＳＧＭＬ文書を処理する際には、まず、文書構造を解析する処理（ＳＧＭＬパーサによる構文解析処理）が必要とされる。文書構造の解析の処理では、対象文書を解析しながら、文書型定義との照らし合わせを行い、対象文書において省略されたタグを復元する処理が主として行われる。実際の文書処理において実行される構文解析処理では、これ以外の処理（例えば、属性の復元やエンティティの展開などの処理）も行われる。

【００２０】図１１に例示したＳＧＭＬ文書１１０を対象文書として、タグ（構造）の復元処理を行うと、図１２に示すようなＳＧＭＬ文書１２０が得られる。図１２に示すＳＧＭＬ文書１２０においては、下線を引いた部分が復元されたタグ（構造）を示している。この対象文書は、図１０に示すような文書型定義１００を参照して、省略されたタグが復元されたもである。つまり、「章」の構造の規定から、タグ＜章＞の次には必ずタグ＜題＞が存在しなければならないので、まず、タグ＜章＞の次にタグ＜題＞を復元し、同様に、タグ＜図＞の次には必ずタグ＜題＞が存在しなければならないので、タグ＜図＞の次にタグ＜題＞を復元する。また、それぞれのエンドタグが省略されているので、内容部分の次に（対応する次のタグの手前の位置に）、それぞれのエンドタグ＜／題＞，＜／段落＞などを復元する。このようにして下線が引かれたようにタグ（構造）が復元される。

【００２１】次に、このようにタグが復元され、構造表現されたＳＧＭＬ文書１２０において、構造を検索する場合の処理について説明する。構造化文書においては、文書編集を行う場合、単なるテキストの文字列の検索だけでなく、文書構造における構造の検索も文書処理の重要な処理となる。これは、構造化文書の処理を行う場合においては、文書構造の構造を利用した編集処理が積極的に行なわれるためである。

【００２２】構造化文書の検索においては、従来のような文字列の検索だけではなく、構造を積極的に利用した検索も有効に利用される。例えば、文書内のＳＧＭＬに関連した図を検索したい場合、これまでの検索処理で

6

は、文書内を全文検索（文字列検索）を行い、テキストの文字列からその「関連した図」の文字列を捜し出していた。

【００２３】しかし、文書構造の構造自体を検索に用いると、例えば、「図のタイトルにＳＧＭＬを含んでいる図」や「図の下部構造の題」のように文書構造における構造を指示して、検索を行うことができ、より対象を絞った検索を行うことができることになる。また、その場合の検索処理も、文書の構造に従って対象範囲が限定して検索できるので、検索処理の効率が良くなるという利点も持つ。

【００２４】前述したように、ＳＧＭＬ文書は、テキスト中にマーク付けを行うだけのタグを埋め込んだ形式の文書アーキテクチャとなっているため、従来からのテキスト処理システムとの親和性が高く、構造がマーク付けのタグで表現されるため、構造を検索する際にも特別な装置あるいは処理プログラムを用いなくとも良い。タグの文字列を検索するという文字列検索を用いて、文書構造の検索を行うことができる。つまり、従来からのテキスト処理装置（文書エディタなど）により、ＳＧＭＬ文書を作成することができ、基本的にはスタートタグとそれに対応するエンドタグを、タグの文字列を検索するという従来の文字列検索のテキスト検索手法を用いて検索することにより、構造検索を行うことができる。

【００２５】

【発明が解決しようとする課題】このように、ＳＧＭＬなどのタグ付き文書の規格が標準化され、多方面で利用され始めている。このタグ付き文書は、フォーマット情報や、その他の文書に関する構造の情報を示すマークアップをタグにより表現する手法を取り入れた文書であるため、ＳＧＭＬの標準化の規格によって、文書の内容が構造化されて、その内容の利用が容易になる。このため、企業、官庁、学校の内外を問わず、この種の文書の流通が盛んになり始めている。このように、タグ付き文書が電子化されて、蓄積されていくことにより、大規模な電子化文書の資源が蓄えられ、利用可能になる。

【００２６】しかし、ＳＭＧＬなどのタグ付き文書の中から、必要な情報を探す場合、文書構造は、タグを利用することにより容易に判定されるので、構造の検索は容易になっているが、文書内容については、これまでのフルテキストサーチなどの手法を利用しなければならず、十分に効率よく内容の検索までは行えないという問題があった。なお、タグを利用することにより、文書構造の位置関係など判別できるので、このようなタグを利用して、タグを検索結果の参照単位の区切りとすることができ、文書内容の利用が効率化できる。

【００２７】本発明は、上述のような問題を解決するためになされたものであり、本発明の目的は、タグ付き文書において、タグを検索結果の参照単位として、効率よく本文中の単語位置を検索し、本文の参照単語を得るこ

とができる全文登録語検索装置および方法を提供することにある。

【００２８】

【課題を解決するための手段】上記のような目的を達成するため、本発明の第１の特徴とする全文登録語検索装置は、タグを有する文書の本文を収めた本文ファイルを入力し、タグで区切られた参照単位に分割する参照単位分割部と、参照単位に含まれる検索対象とする単語に対して、単語と当該単語が出現する本文における参照単位の位置の対を抽出する単語位置抽出部と、抽出された単語と参照単位の位置の対を単語に従って分類し、単語に対し当該単語が出現する全ての参照単位の位置を組とした単語位置集合を得る分類部と、単語位置集合に対し、単語から位置集合を得る検索インデックスを生成する検索インデックス生成部とを備えることを特徴とする。

【００２９】また、本発明において、第２の特徴とする全文登録語検索方法は、タグを有する文書の本文を収めた本文ファイルをタグで区切られた参照単位に分割し、参照単位内に含まれる検索対象とする単語に対して、単語と当該単語が出現する全ての参照単位の位置の対を抽出し、抽出され単語と参照単位の位置の対を単語により分類し、単語と当該単語が出現する全ての参照単位の位置を組とした単語位置集合を作成し、作成された単語位置集合に基づいて、単語から位置集合を得ることができる検索インデックスを生成することを特徴とする。

【００３０】また、本発明の第３の特徴とする全文登録語検索装置では、更に、検索インデックス生成部により作成された検索インデックスを用いて得られたタグの位置の集合に基づいて、参照単位を次のタグまであるいは適当な長さだけ表示して、検索結果を表示する検索処理部を有することを特徴とする。

【００３１】また、本発明の第４の特徴とする全文登録語検索装置は、タグに付加されたフィールド情報に基づいて特定のフィールドを検索対象とする場合、本文中の単語について、その単語の直前にフィールドの種別を示す文字列を付加したものを新たな単語とし、当該単語が出現する位置の直前にあるタグの位置との対を単語位置集合対とし、単語を入力し、その語が出現する位置の直前にあるタグの位置の集合を結果として出力することを特徴とする。

【００３２】また、本発明の第５の特徴とする全文登録語検索装置は、本文中に指定形式で記述された属性と属性の値の対を含む場合、属性と値の対を単語として登録し、検索時に指定形式の一部を入力することにより、検索対象の指定形式の文字列から始まる指定形式の属性と値との対を列挙することを特徴とする。

【００３３】また、本発明の第６の特徴とする全文登録語検索装置は、特定のフィールドを検索対象とし、本文中に指定形式で記述された属性と値の対を含む場合、属性と値の対の直前にフィールドの種別を示す文字列を付

加したものを単語として登録し、検索時に指定形式の一部を入力することにより、検索対象の指定形式の文字列から始まる指定形式を列挙することを特徴とする。

【００３４】

【作用】本発明の第１の特徴とする全文登録語検索装置においては、タグを有する文書に対して、まず、参照単位分割部が、タグを有する文書の本文を収めた本文ファイルを入力し、タグで区切られた参照単位に分割すると、単語位置抽出部が、参照単位に含まれる検索対象とする単語に対して、単語と当該単語が出現する本文における参照単位の位置の対を抽出する。次に、分類部が、抽出された単語と参照単位の位置の対を単語に従って分類し、単語に対し当該単語が出現する全ての参照単位の位置を組とした単語位置集合を得る。そして、検索インデックス生成部が、単語位置集合に対し、単語から位置集合を得る検索インデックスを生成する。これにより、文書内の全ての単語は、その直前のタグの位置と共に、検索インデックスに登録されるので、検索インデックスを用いて検索を行うことによって、文書中の単語の参照単位のタグ位置が直ちに検索でき、高速に参照単位の本文部分が表示出力されることになる。

【００３５】また、本発明の第２の特徴とする全文登録語検索方法においては、検索対象の文書に対して、検索処理を開始する前に、まず、タグを有する文書の本文を収めた本文ファイルをタグで区切られた参照単位に分割し、参照単位内に含まれる検索対象とする単語に対して、単語と当該単語が出現する全ての参照単位の位置の対を抽出し、抽出され単語と参照単位の位置の対を単語により分類し、単語と当該単語が出現する全ての参照単位の位置を組とした単語位置集合を作成し、作成された単語位置集合に基づいて、単語から位置集合を得ることができる検索インデックスを生成する。これにより、検索インデックスクスが作成された後は、次回の検索処理から、作成された検索インデックスを利用することにより、直ちに検索対象の単語の参照単位のタグ位置が直ちに検索でき、高速に参照単位の文書部分が表示出力されることになる。

【００３６】また、本発明の第３の特徴とする全文登録語検索装置においては、検索処理部が、検索インデックス生成部により作成された検索インデックスを用いて得られたタグの位置の集合に基づいて、参照単位を次のタグまであるいは適当な長さだけ表示して、検索結果を表示する。すなわち、検索対象の単語を検索する場合、検索インデックス生成部により作成された検索インデックスを利用することにより、検索結果として、単語に対してタグの位置の集合が得られるので、これにより、得られたタグの位置の集合に基づいて、参照単位を次のタグまであるいは適当な長さだけ表示して、検索結果を表示する。このため、検索結果の対象文書の箇所の表示が能率よく行える。

【００３７】また、本発明の第４の特徴とする全文登録語検索装置は、タグに付加されたフィールド情報に基づいて特定のフィールドを検索対象とする場合、本文中の単語について、その単語の直前にフィールドの種別を示す文字列を付加したものを新たな単語とし、当該単語が出現する位置の直前にあるタグの位置との対を単語位置集合対とする。そして、検索する単語を入力し、その単語が出現する位置の直前にあるタグの位置の集合を結果として出力する。これにより、タグに付加されたフィールド情報に基づいて特定のフィールドを検索対象とすることができる。

【００３８】また、本発明の第５の特徴とする全文登録語検索装置においては、更に、本文中に指定形式で記述された属性と属性の値の対を含む場合、属性と値の対を単語として登録し、検索時に指定形式の一部を入力することにより、検索対象の指定形式の文字列から始まる指定形式の属性と値の対を列挙する。これによって、本文の中において、書式など指定形式で記述された属性と属性の値の対を含む場合においても、それを検索対象とした検索が可能となる。

【００３９】本発明の第６の特徴とする全文登録語検索装置においては、特定のフィールドを検索対象とし、本文中に指定形式で記述された属性と値の対を含む場合、属性と値の対の直前にフィールドの種別を示す文字列を付加したものを単語として登録し、検索時に指定形式の一部を入力する。これによって、検索対象の指定形式の文字列から始まる指定形式を列挙する。このため、特定のフィールドを検索対象とし、本文中に指定形式で記述された属性と値の検索に対しても容易に対応できる。

【００４０】このようにして、本発明の全文登録語検索装置によれば、本文中の検索対象とする単語について、例えば、本文中の全ての単語について、その単語を出現する位置の直前にあるタグの位置を全て集めて、検索インデックスを作成する。この検索インデックスを用いることにより、検索処理を行う場合、検索対象の単語を入力とし、その単語が出現する位置の直前にあるタグの位置の集合が、検索結果として出力できる。つまり、タグを有する文書において、タグを検索結果の参照単位の区切りとすると、文書内に表われる全ての単語が、その直前のタグの位置と共に、検索インデックスとして保存される。このため、文書内の全ての単語に対し、当該単語を含むタグで区切られた参照単位を即座に検索することが可能となる。

【００４１】

【実施例】以下、本発明の一実施例を図面を用いて具体的に説明する。図１は本発明の実施例の全文登録語検索装置の装置構成の要部を示すブロック図である。図１において、１は検索語登録処理部、２は登録語検索処理部、３はタグを有する文書の本文ファイル、４は本文ファイルに付加された検索インデックス、１１は参照単位分割部、１２は単語位置抽出部、１３は分類処理機構、１４は検索インデックス生成部、１５は検索単語入力部、１６は検索機構、１７は本文参照単位出力部、１８は検索結果出力表示部である。

【００４２】ここでの全文登録語検索装置においては、検索対象の単語を入力して本文検索の処理を実行する前に、その前処理として、検索語登録処理部１が、タグを有する文書の本文ファイル３から、検索対象とする単語の登録を行い、本文ファイルに付加する検索インデックス４を作成する。検索インデックス４が付加された本文ファイルは、登録語検索処理部２において、本文ファイルに付加された検索インデックス４を利用して、検索対象の本文からその登録語を検索する処理が行われる。図１を参照して説明する。

【００４３】検索語登録処理部１において、まず、参照単位分割部１１にタグを有する文書の本文ファイル３を入力する。参照単位分割部１１は、本文ファイル３が入力されると、本文をタグで区切られた参照単位に分割する。この参照単位の本文を入力として、次に、単語位置抽出部１２が、検索対象とする単語として、単語と当該単語が出現する本文における参照単位の位置の対を抽出する。

【００４４】次に、分類処理機構１３が、抽出された単語と参照単位の位置の対を単語に従って分類し、後述するように、単語に対し当該単語が出現する全ての参照単位の位置を組とした単語位置集合を生成する。そして、検索インデックス生成部１４が、得られた単語位置集合に対し、単語から位置集合を得る検索インデックス４を生成し、本文ファイル３に対して、その対応する検索インデックス４を付加して、検索インデックス４付きの本文ファイルを作成する。これにより、ここでの検索インデックス４が付加された本文ファイルは、検索インデックス４を用いる登録語検索処理部２の検索処理により、検索対象の単語から高速に本文ファイルのタグで区切られた参照単位の位置を得ることができ、該当の参照単位の内容を直ちに表示できる。

【００４５】検索対象の単語から本文ファイルの検索を行う場合、登録語検索処理部２においては、検索単語入力部１５を介して、検索対象とする単語を入力すると、検索機構１６が、本文ファイルに付加された検索インデックス４を用いて、検索対象の単語の検索処理を行い、その単語の対応の参照単位の位置の集合を検索する。参照単位の位置が検索できると、次に、本文参照単位出力部１７が、その参照単位の位置から本文ファイルをアクセスして、該当の参照単位を直ちに出力し、検索結果出力表示部１８を介して、該当の参照単位を出力表示する。

【００４６】このように、タグを有する文書の本文ファイル３を、検索語登録処理部１の参照単位分割部１１に入力すると、参照単位分割部１１は、タグで区切られた

参照単位に分割し、参照単位分割部１１から参照単位とその位置を得る。単語位置抽出部１２は、この参照単位を入力とし、この参照単位内に含まれる全ての単語について、その単語と、当該単語が出現する本文における参照単位の位置の対を生成する。次に、分類処理機構１３が、それぞれの単語について、その単語が表れる全ての参照単位に位置の組である（単語・参照単位の位置集合）対を得る。次に、検索インデックス生成部１４により、全ての（単語・参照単位の位置集合）対から、各々の単語についての参照単位の位置集合を生成し、検索インデックス４を作成する。

【００４７】これにより、文書内の全ての単語は、その直前のタグ（参照単位）の位置の情報と共に、検索インデックス４に登録されるので、検索対象とする単語から検索インデックス４を用いて検索することにより、文書中の単語の参照単位のタグ位置が直ちに検索でき、高速に参照単位の文書内容の部分が表示出力される。

【００４８】図２は、第１の実施例の本文ファイルに対する検索インデックスの作成処理を示す処理フローを示すＰＡＤ（Problem Analysis Diagram）図である。また、図３は、図２に示す処理フローにより検索インデックスを作成する場合の作成プロセスの要部を具体的に説明する図である。図２および図３を参照して、全文登録語検索のための検索インデックスの作成処理を説明する。

【００４９】まず、図２を参照して、本文ファイルに対する検索インデックスの作成処理の処理フローを概要を説明する。処理を開始すると、処理ブロック２１において、本文ファイル３１をタグの位置で分割し、分割した部分を参照単位とし、そのタグ位置を一時記憶する処理を行う。次に、繰り返し処理の制御ブロック２２の処理を行う。この制御ブロック２２の処理では、タグで分割された全てのタグ位置とその本文部分の対に対して、次の処理ブロック２３および処理ブロック２４の処理を繰り返し行う処理制御を行う。

【００５０】この制御ブロック２２の制御下の繰り返し処理では、まず、処理ブロック２３において、処理対象の参照単位のタグ位置をＡファイルに書き出す。次に、処理ブロック２４において、処理対象の参照単位の本文の単語を、Ａファイルに先に書き出したタグ位置に続いて、順番にＡファイルに書き出す。これにより、Ａファイルには、１つの参照単位について、タグ位置に続いて、その本文中の単語が連続して書き出される。このような処理を全ての参照単位について、制御ブロック２２の処理制御により、繰り返し行う。このため、Ａファイルには、図３に示すように、本文ファイル３１から各々の参照単位について、まず、タグ位置が書き出されて、続いて当該タグ位置に対応する参照単位の本文の中の単語が順次に書き出される。この結果、Ａファイル３２の内容は、タグ位置とそれに続く単語の組３２ａが、参照

単位の数だけ続くデータが得られる。

【００５１】このようにして、Ａファイルが作成されると、次に、繰り返し処理の制御ブロック２５の処理を行う。この制御ブロック２５の処理では、Ａファイルに含まれる全ての単語に対して、次の処理ブロック２６および処理ブロック２７の処理を繰り返し行う処理制御を行う。

【００５２】この制御ブロック２５の制御下の繰り返し処理では、まず、処理ブロック２６において、単語をキーとして、当該単語に対応するタグ位置を値とする対を作成する。続いて、処理ブロック２７において、同じキー（単語）を持つ値（タグ位置）の対を集めて、キーと値の集合から構成されるリストを作成し、これをＢファイルに書き出す。これにより、Ｂファイルには、１つの単語について、その単語が出現する参照単位のタグ位置のリストが得られる。このような処理を全てのＡファイルの単語について、制御ブロック２５による処理制御により、繰り返し行う。

【００５３】この結果、図３に示すように、Ｂファイル３３には、本文ファイルの各タグに区切られる参照単位の全ての単語について、当該単語がその出現する各々の参照単位に対応するタグ位置のリストが得られる。図３に示すＢファイル３３の例で説明すると、第１番目の単語１および第２番目の単語２に対応して、それぞれに『（単語１，０，…）』および『（単語２，０，１００，…）』のリストデータが得られている。つまり、これらのリストデータは、それぞれに『単語１が出現する参照単位のタグ位置がアドレス“０”，…であること』および『単語２が出現する参照単位のタグ位置がアドレス“０”，アドレス“１００”，…であること』を意味している。

【００５４】次に、処理ブロック２８の処理を行い、Ｂファイルの内容に基づいて、単語からタグ位置の集合を検索できる検索インデックスを作成し、ここでの処理を終了する。これにより、各々の単語に対する検索インデックスが作成されると、その検索インデックスを用いることにより、検索対象の単語から直ちに、その単語が出現する参照単位のタグ位置の集合が得られる。したがって、検索単語から得られたタグ位置の集合に従って、当該タグ位置の集合からそれぞれの参照単位を表示できる。

【００５５】以上に説明した全文登録語検索装置の第１の実施例においては、本文ファイルの参照単位をタグにより区分し、その位置を指示するタグ位置と、その中に含まれる単語を求めて記録する場合（Ａファイル）、最初に参照単位の開始を指示すタグ位置を置き、続いて、その参照単位に属する単語を書くファイル形式をとっているが、各々の単語と参照単位（タグのタイプ）の間の関係を明確にして、同じ種類のタグの参照単位の中の単語を他と区別するため、各々の単語のデータにタグの種

類を示すフィールドを設けるようにしてもよい。これにより、同じ種類のタグの参照単位を検索単位として扱える。このような例を第２の実施例として説明する。

【００５６】図４は、第２の実施例の全文登録語検索装置の本文ファイルに対する検索インデックスの作成処理を示す処理フローを示すＰＡＤ図である。また、図５は、図４に示す処理フローによる検索インデックスの作成プロセスの要部を具体的に説明する図である。図４および図５を参照して、第２の実施例の全文登録語検索のための検索インデックスの作成処理を説明する。

【００５７】第２の実施例においては、単語から検索された結果のタグ位置により、表示する参照単位の区切りのタグの種類が直ちに判別できるように、参照単位ごとにその本文の検索対象となる単語の前にタグの種類を示すフィールドの文字列を付加している。これは、例えば、単語位置抽出部において、抽出した単語の前にタグ種別を示すフィールドの文字列を付加する処理を追加するように変形することにより、容易に対応できる。この種のタグの種類を示すフィールドを用いる場合の一例として、例えば、本文ファイルの参照単位の内容が、故障の個々の内容を示している場合に、故障の症状，原因，対処の３つのフィールドを１レコード中に設けておき、そのようなレコードが繰返し現われるような文書において、その単語の検索範囲を、症状を示すフィールドだけに限定する場合などに利用できる。この場合、タグの種類を示すフィールドの文字列として、症状，原因，対処の３つの種類を示す文字列を付加する。

【００５８】図４を参照して、本文ファイルに対する検索インデックスの作成処理の処理フローを概要を説明する。処理を開始すると、処理ブロック４１において、本文ファイルをタグの位置で分割し、分割した部分を参照単位とし、そのタグ位置を一時記憶する処理を行う。次に、繰り返し処理の制御ブロック４２の処理を行う。この制御ブロック４２の処理では、タグで分割された全てのタグ位置とその本文部分の対に対して、次の処理ブロック４３および処理ブロック４４の処理を繰り返し行う処理制御を行う。

【００５９】この制御ブロック４２の制御下の繰り返し処理では、まず、処理ブロック４３において、処理対象の参照単位のタグ位置をＣファイルに書き出す。次に、処理ブロック４４において、処理対象の参照単位の本文の各々の単語に対して、本文の単語の前にタグの種類を示すフィールドの文字列を付加したものを、新たな単語として、Ｃファイルに先に書き出したタグ位置に続いて、順番にＣファイルに書き出す。これにより、Ｃファイルには、１つの参照単位について、タグ位置に続いて、タグの種類を示すフィールドの文字列を付加した本文中の単語が連続して書き出される。

【００６０】このような処理を全ての参照単位について、制御ブロック４２の処理制御により、繰り返し行う。この結果、図５に示すように、Ｃファイル５２には、本文ファイル５１から各々の参照単位について、まず、タグ位置が書き出されて、続いて当該タグ位置に対応する参照単位のタグの種類を示すフィールドの文字列（fld１など）を前に付加した本文の中の単語（単語１，単語２など）が順次に書き出される。この結果、Ｃファイル５２の内容として、タグ位置とそれに続くタグの種類を示すフィールドの文字列を付加した単語の組５２ａが、参照単位の数だけ続くデータが得られる。

【００６１】このようにして、Ｃファイルが作成されると、次に、繰り返し処理の制御ブロック４５の処理を行う。制御ブロック４５の処理では、Ｃファイルに含まれる全ての単語に対して、次の処理ブロック４６および処理ブロック４７の処理を繰り返し行う処理制御を行う。

【００６２】この制御ブロック４５の制御下の繰り返し処理では、まず、処理ブロック４６において、単語をキーとして、当該単語に対応するタグ位置を値とする対を作成する。続いて、処理ブロック４７において、同じキー（単語）を持つ値（タグ位置）の対を集めて、キーと値の集合から構成されるリストを作成し、これをＤファイルに書き出す。これにより、Ｄファイルには、前にタグの種類を示すフィールドの文字列を付加した１つの単語について、その単語が出現する参照単位のタグ位置のリストが得られる。このような処理を全てのＣファイルに書き出された単語について、制御ブロック４５による処理制御により、繰り返し行う。

【００６３】この結果、図５に示すように、Ｄファイル５３には、本文ファイル５１の全ての単語について、前にタグの種類を示すフィールドの文字列が付加された単語毎に、当該単語がその出現する個々の参照単位に対応して、そのタグ位置のリストが得られる。図５に示す例で説明すると、第１番目の単語１および第２番目の単語２に対しては、それぞれ『（fld１-単語１，０，…）』および『（fld１-単語２，１００，…）』のリストデータが得られている。つまり、これらのリストデータは、『タグの種類が“fld１”である参照単位で単語１が出現するタグ位置が、アドレス“０”，…であること』および『タグの種類が“fld１”である参照単位で単語２が出現するタグ位置がアドレス“０”，アドレス“１００”，…であること』をそれぞれ意味している。

【００６４】次に、処理ブロック４８の処理を行い、Ｄファイルの内容に基づいて、単語からタグ位置（タグフィールド名）の集合を検索できる検索インデックスを作成して、ここでの処理を終了する。これにより、各々の単語に対する検索インデックスが作成されると、その検索インデックスを用いることにより、タグの種類と検索対象の単語を指定することにより、タグの種類に応じて異なる参照単位についての検索対象の単語から直ちに、その単語が出現する参照単位のタグ位置の集合が得られる。したがって、検索単語から得られたタグ位置の集合

に従って、当該タグ位置からそれぞれの参照単位を表示できる。

【００６５】以上に説明した全文登録語検索装置の第２の実施例においては、本文ファイルの参照単位をタグにより区分し、タグの種類に応じて、その位置を指示するタグ位置と、その中に含まれる単語を求めて記録する場合（Ｃファイル）、最初に参照単位の開始を指示すタグ位置を置き、続いて、その参照単位に属する単語に対しては、タグの種類を示すフィールドの文字列を付加して書くファイル形式をとっている。これにより、タグの種類に応じて、それぞれの参照単位の中の検索対象の単語の検索を、タグで区切る参照単位毎に高速に行うことができる。

【００６６】また、タグを検索結果の参照単位の区切りとする全文登録語検索を行う場合において、本文の検索対象とする単語の中で、属性と値の対の記述が存在するものについては、その属性と値の対を検索対象の登録語として、登録しておくことにより、これらの単語の属性と値の対を検索対象して、前述の実施例と同様に、高速にタグの区切りを参照単位とする検索を行うことができる。このような実施例を、第３の実施例として説明する。

【００６７】第３の実施例の全文登録語検索装置においては、タグを検索結果の参照単位の区切りとする全文登録語検索を行う場合、本文ファイルの本文の単語に、属性と値の対の記述のあるものについては、その対の記述を単語として登録する。具体的に説明すると、ここでの属性と値の対の例としては、本文中に、例えば、｛売上高＝１００００｝などのように、特別な形式により単語が示され、その単語が他と区別されている場合などがある。

【００６８】図６は、第３の実施例の全文登録語検索装置の本文ファイルに対する検索インデックスの作成処理を示す処理フローを示すＰＡＤ図である。また、図７は、図６に示す処理フローにより検索インデックスを作成する作成プロセスの要部を具体的に説明する図である。図６および図７を参照して、第３の実施例の全文登録語検索のための検索インデックスの作成処理を説明する。

【００６９】まず、図６を参照して、第３の実施例の本文ファイルに対する検索インデックスの作成処理の処理フローを概要を説明する。処理を開始すると、処理ブロック６１において、本文ファイルをタグの位置で分割し、分割した部分を参照単位とし、そのタグ位置を一時記憶する処理を行う。次に、繰り返し処理の制御ブロック６２の処理を行う。この制御ブロック６２の処理では、タグで分割された全てのタグ位置とその本文部分（参照単位）の対に対して、次の処理ブロック６３および処理ブロック６４の処理を繰り返し行う処理制御を行う。

【００７０】制御ブロック６２の制御下の繰り返し処理では、まず、処理ブロック６３において、処理対象の参照単位のタグ位置をＥファイルに書き出す。次に、処理ブロック６４において、処理対象の参照単位の本文の単語を、Ｅファイルに先に書き出したタグ位置に続いて、順番にＥファイルに書き出す。ただし、属性と値の対の記述のあるものについては、その対も単語として、順番にＥファイルに書き出す。これにより、Ｅファイルには、１つの参照単位について、当該タグ位置に続いて、その本文中の単語と、あれば属性と値の対とが連続して書き出される。このような処理を全ての参照単位について、制御ブロック６２の処理制御により、繰り返し行う。

【００７１】この結果、図７に示すように、Ｅファイル７２には、本文ファイル７１から各々の参照単位について、まず、タグ位置が書き出されて、続いて当該タグ位置に対応する参照単位の本文の中の単語と、属性と値の対とが順次に書き出される。このため、前述の場合と同様に、Ｅファイル７２の内容は、タグ位置とそれに続く単語と、属性と値の対との組が、参照単位の数だけ続くデータが得られる。

【００７２】このようにして、Ｅファイルが作成されると、次に、繰り返し処理の制御ブロック６５の処理を行う。制御ブロック６５の処理では、Ｅファイルに含まれる全ての単語（属性と値の対を含む）に対して、次の処理ブロック６６および処理ブロック６７の処理を繰り返し行う処理制御を行う。

【００７３】この制御ブロック６５の制御下の繰り返し処理では、まず、処理ブロック６５において、単語をキーとして、当該単語に対応するタグ位置を値とする対を作成する。続いて、処理ブロック６７において、同じキー（単語）を持つ値（タグ位置）の対を集めて、キーと値の集合から構成されるリストを作成し、これをＦファイルに書き出す。これにより、Ｆファイルには、１つの単語について、その単語が出現する参照単位のタグ位置のリストが得られる。このような処理を全てのＥファイルの単語（属性と値の対を含む）について、制御ブロック６５による処理制御により、繰り返し行う。

【００７４】この結果、Ｆファイル７３には、図７に示すように、本文ファイル７１における全ての単語（属性と値の対を含む）について、当該単語がその出現する各々の参照単位に対応するタグ位置のリストが得られる。ここで、図７に示すＦファイル７３の例で説明すると、第１番目の単語１および第２番目の単語２に対応して、それぞれに『（単語１，０，…）』および『（単語２，０，１００，２２０，…）』のリストデータが得られている。つまり、これらのリストデータは、それぞれに『単語１が出現する参照単位のタグ位置がアドレス“０”，…であること』および『単語２が出現する参照単位のタグ位置がアドレス“０”，アドレス“１０

０”，アドレス“２２０”，…であること』を意味している。また、ここでは、属性と値の対についても、単語の場合と同様に『（｛××事業規模＝１０００｝，０，…）』のリストデータが得られており、このリストデータは、『属性と値の対｛××事業規模＝１０００｝が出現する参照単位のタグ位置がアドレス“０”…であること』を意味している。

【００７５】次に、処理ブロック６８の処理を行い、Ｆファイルの内容に基づいて、単語からタグ位置の集合を検索できる検索インデックスを作成し、ここでの処理を終了する。これにより、各々の単語に対する検索インデックスが作成されると、その検索インデックスを用いることによって、検索対象として単語を指示することにより、前述の場合と同様に、検索対象の単語が出現する参照単位のタグ位置の集合が得られる。また、例えば、検索対象として属性と値の対を指示することにより、直ちに、検索対象の属性と値の対が出現する参照単位のタグ位置の集合が得られる。したがって、検索単語から得られたタグ位置の集合に従って、当該タグ位置の集合からそれぞれの参照単位を表示できる。

【００７６】次に、前述した第２の実施例による単語の前にタグの種類を示すフィールドの文字列を付加し、更に、第３の実施例にによる属性と値の対のあるものについては、その属性と値の対を単語として登録する場合の変形を組合せるようにしても良い。このような実施例について、第４の実施例として説明する。

【００７７】つまり、第４の実施例による全文登録語検索装置は、タグを検索結果の参照単位の区切りとする全文登録語検索装置のうち、属性と値の対の記述のあるものについては、その対を単語として登録し、その際、単語として登録する属性と値の対の前にタグの種類を示すフィールドの文字列を付加して登録する。

【００７８】図８は、第４の実施例の全文登録語検索装置の本文ファイルに対する検索インデックスの作成処理を示す処理フローを示すＰＡＤ図である。また、図９は、図８に示す処理フローによる検索インデックスの作成プロセスの要部を具体的に説明する図である。図８および図９を参照して、第４の実施例の全文登録語検索のための検索インデックスの作成処理を説明する。

【００７９】第４の実施例においては、単語または属性と値の対の指定により、検索された結果のタグ位置により、表示する参照単位の区切りのタグの種類が直ちに判別できるように、参照単位ごとにその本文の検索対象となる単語（属性と値の対を含む）の前にタグの種類を示すフィールドの文字列を付加する。これも、前述したように、例えば、単語位置抽出部において、抽出した単語の前にタグ種別を示すフィールドの文字列を付加する処理を追加するように変形することにより、容易に対応できる。

【００８０】図８を参照して、本文ファイルに対する検索インデックスの作成処理の処理フローを概要を説明する。処理を開始すると、処理ブロック８１において、本文ファイルをタグの位置で分割し、分割した部分を参照単位とし、そのタグ位置を一時記憶する処理を行う。次に、繰り返し処理の制御ブロック８２の処理を行う。この制御ブロック８２の処理では、タグで分割された全てのタグ位置とその本文部分の対に対して、次の処理ブロック８３および処理ブロック８４の処理を繰り返し行う処理制御を行う。

【００８１】この制御ブロック８２の制御下の繰り返し処理では、まず、処理ブロック８３において、処理対象の参照単位のタグ位置をＧファイルに書き出す。次に、処理ブロック８４において、処理対象の参照単位の本文の各々の単語に対して、本文の単語の前にタグの種類を示すフィールドの文字列を付加したものを、新たな単語として、Ｇファイルに先に書き出したタグ位置に続いて、順番にＧファイルに書き出す。ただし、この場合、属性と値の対の記述のあるものについても、その対を単語して、順番にＧファイルに書き出す。これにより、Ｇファイルには、１つの参照単位について、タグ位置に続いて、タグの種類を示すフィールドの文字列を前に付加した本文中の単語または属性と値の対が連続して書き出される。

【００８２】このような処理を全ての参照単位について、制御ブロック８２の処理制御により、繰り返し行う。この結果、図９に示すように、Ｇファイル９２には、本文ファイル９１から各々の参照単位について、まず、タグ位置が書き出されて、続いて当該タグ位置に対応する参照単位のタグの種類を示すフィールドの文字列を前に付加した本文の中の単語が順次に書き出され、または、タグの種類を示すフィールドの文字列を付加した本文の中の属性と値の対が書き出される。この結果、Ｇファイル９２の内容として、タグ位置とそれに続くタグの種類を示すフィールドの文字列を付加した単語または属性と値の対の組が、参照単位の数だけ続くデータが得られる。

【００８３】このようにして、Ｇファイルが作成されると、次に、繰り返し処理の制御ブロック８５の処理を行う。制御ブロック８５の処理では、Ｇファイルに含まれる全ての単語（属性と値の対を含む）に対して、次の処理ブロック８６および処理ブロック８７の処理を繰り返し行う処理制御を行う。

【００８４】この制御ブロック８５の制御下の繰り返し処理では、まず、処理ブロック８６において、単語（属性と値の対を含む）をキーとして、当該単語に対応するタグ位置を値とする対を作成する。続いて、処理ブロック８７において、同じキー（単語，属性と値の対）を持つ値（タグ位置）の対を集めて、キーと値の集合から構成されるリストを作成し、これをＨファイルに書き出す。これにより、Ｈファイルには、タグの種類を示すフ

19

ィールドの文字列を前に付加した１つの単語について、その単語が出現する参照単位のタグ位置のリストが得られる。このような処理を全てのＧファイルに書き出されている単語（属性と値の対を含む）について、制御ブロック８５による処理制御により、繰り返し行う。

【００８５】この結果、Ｈファイル９３には、図９に示すように、本文ファイル９１の全ての単語について、タグの種類を示すフィールドの文字列を前に付加した１つの単語毎に、当該単語がその出現する個々の参照単位に対応して、そのタグ位置のリストが得られる。図９に示すＨファイル９３の例で説明すると、第１番目の単語１および第２番目の単語２に対しては、それぞれ『（fld１-単語１，０，…）』および『（fld１-単語２，０，２２０，…）』のリストデータが得られている。つまり、これらのリストデータは、『タグの種類が“fld１”である参照単位で単語１が出現するタグ位置が、アドレス“０”，…であること』および『タグの種類が“fld１”である参照単位で単語２が出現するタグ位置がアドレス“０”，アドレス“２２０”，…であること』をそれぞれ意味している。また、属性と値の対についても、単語の場合と同様な形式で『（fld１-｛××事業規模＝１０００｝，０，…）』のリストデータが得られており、このリストデータは、『タグの種類が“fld１”である参照単位で、属性と値の対｛××事業規模＝１０００｝が出現するタグ位置がアドレス“０”…であること』を意味している。

【００８６】次に、処理ブロック８８の処理を行い、Ｈファイルの内容に基づいて、単語からタグ位置（タグフィールド名）の集合を検索できる検索インデックスを作成して、ここでの処理を終了する。これにより、各々の単語に対する検索インデックスが作成されると、その検索インデックスを用いることにより、タグの種類と検索対象の単語として属性と値の組を指定することにより、タグの種類に応じて異なる参照単位についての検索対象の属性の値の組から直ちに、その属性と値の組が出現する参照単位のタグ位置の集合が得られる。したがって、検索単語から得られたタグ位置の集合に従って、当該タグ位置からそれぞれの参照単位を表示できる。

【００８７】以上、本発明の実施例について説明したが、本発明は、上述した実施例に限定されるものではない。ここでは、全ての単語を抽出して登録語とする例について示しているが、助詞、助動詞、接続詞など、検索対象を特定して検索語として登録するようにしてもよいことは明らかである。その際、検索対象の特徴を必ずしも示していないような単語を登録しないことも可能である。

【００８８】

【発明の効果】以上、説明したように、本発明の全文登録語検索装置によれば、本文中の検索対象とする単語に

20

ついて、例えば、本文中の全ての単語について、その単語を出現する位置の直前にあるタグの位置を全て集めて、検索インデックスを作成して検索装置を構成する。そして、検索対象の単語を入力とし、その単語が出現する位置の直前にあるタグの位置の集合を検索結果として出力する。これにより、タグを有する文書において、タグを検索結果の参照単位の区切りとして、文書内の全ての単語が、その直前のタグの位置と共に、検索インデックスとして保存されるので、文書内の全ての単語に対し、当該単語を含むタグで区切られた参照単位を即座に検索することが可能となる。

【図面の簡単な説明】

【図１】　図１は本発明の実施例の全文登録語検索装置の装置構成の要部を示すブロック図、

【図２】　図２は第１の実施例の本文ファイルに対する検索インデックスの作成処理を示す処理フローを示すＰＡＤ（Problem Analysis Diagram）図、

【図３】　図３は図２に示す処理フローにより検索インデックスを作成する場合の作成プロセスの要部を具体的に説明する図、

【図４】　図４は第２の実施例の全文登録語検索装置の本文ファイルに対する検索インデックスの作成処理を示す処理フローを示すＰＡＤ図、

【図５】　図５は図４に示す処理フローによる検索インデックスの作成プロセスの要部を具体的に説明する図、

【図６】　図６は第３の実施例の全文登録語検索装置の本文ファイルに対する検索インデックスの作成処理を示す処理フローを示すＰＡＤ図、

【図７】　図７は図６に示す処理フローによる検索インデックスの作成プロセスの要部を具体的に説明する図、

【図８】　図８は第４の実施例の全文登録語検索装置の本文ファイルに対する検索インデックスの作成処理を示す処理フローを示すＰＡＤ図、

【図９】　図９は図８に示す処理フローによる検索インデックスの作成プロセスの要部を具体的に説明する図、

【図１０】　図１０はＳＧＭＬの文書型定義（ＤＴＤ）の一例を示す図、

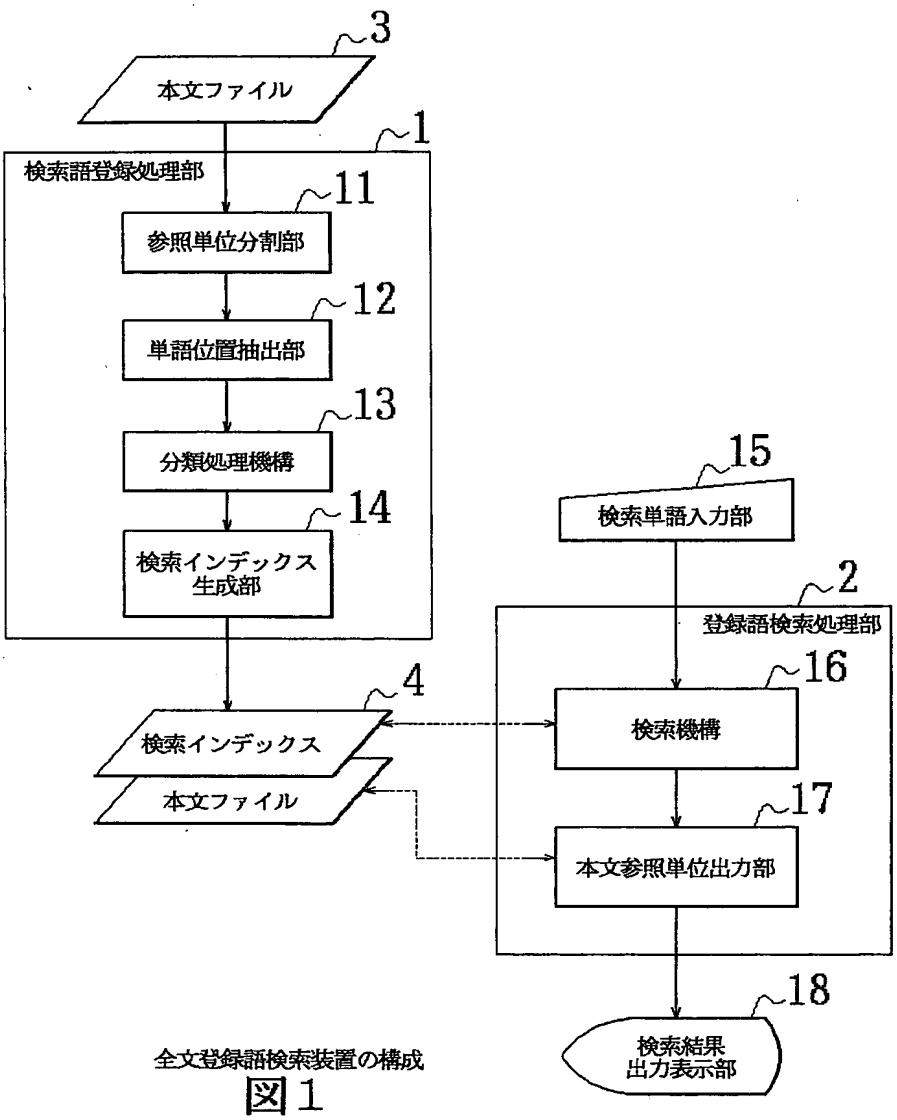
【図１１】　図１１はタグ付き文書としてのタグが省略されたＳＧＭＬ文書の一例を説明する図、

【図１２】　図１２は省略されたタグが復元されたＳＧＭＬ文書の一例を説明する図である。

【符号の説明】

１…検索語登録処理部、２…登録語検索処理部、３…タグを有する文書の本文ファイル、４…本文ファイルに付加された検索インデックス、１１…参照単位分割部、１２…単語位置抽出部、１３…分類処理機構、１４…検索インデックス生成部、１５…検索単語入力部、１６…検索機構、１７…本文参照単位出力部、１８…検索結果出力表示部。

【図１】



全文登録語検索装置の構成

図１

【図１０】

100

```
<!ELEMENT   文書    - -   章*                    >
<!ELEMENT   章      - -   (題, (段落 | 図)*, 章* ) >
<!ELEMENT   題      O O   (#PCDATA)              >
<!ELEMENT   段落    - O   (#PCDATA)              >
<!ELEMENT   図      - O   (題, 図本体)           >
<!ELEMENT   図本体  - O   EMPTY                  >
```

SGMLの文書型定義(DTD)の一例

図１０

【図１１】

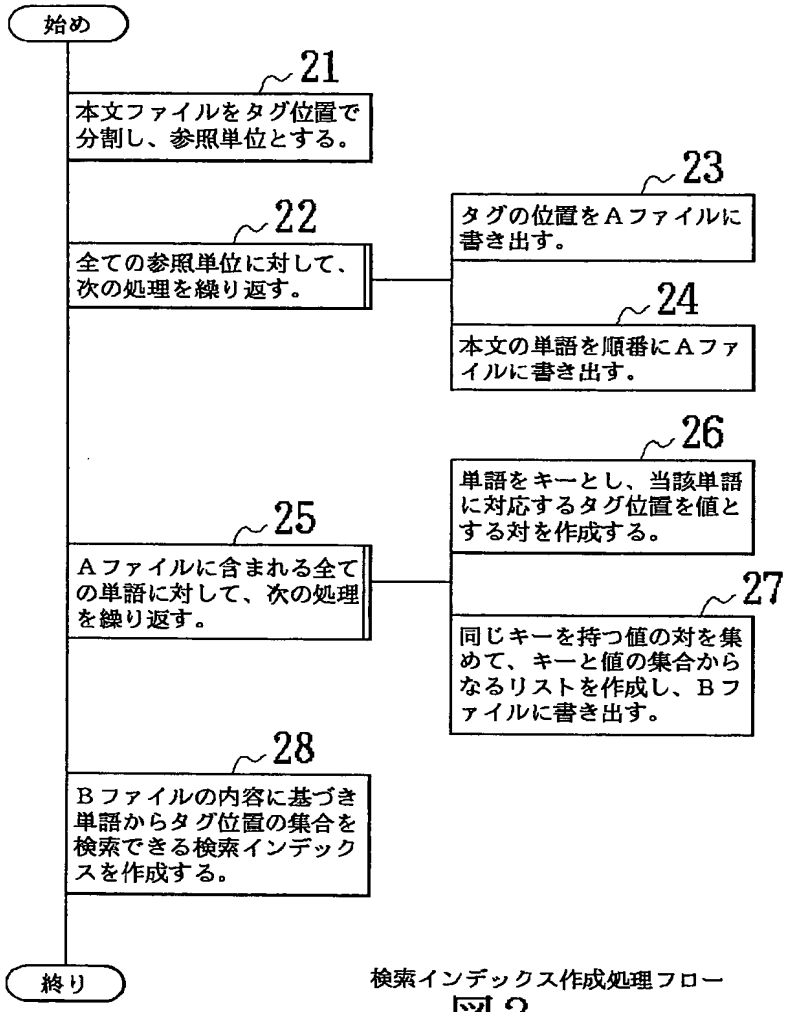
110

```
<文書>
 <章>SGMLとは
  <段落>SGMLとは文書の表現形態を規定した１９８６年に制定
        された国際規格であり、ＩＳＯ８８７９と呼ばれる。
  <段落>SGMLは最近になり、アメリカを中心として普及してき
        ている。
  <章>SGMLの文書表現
   <段落>SGMLはスタートタグとエンドタグと呼ばれるタグを
         テキストに埋め込むことで文書構造を表現する。
    <図>SGML文書例<図本体>
  </章>
 </章>
</文書>
```

SGMLによる文書の例

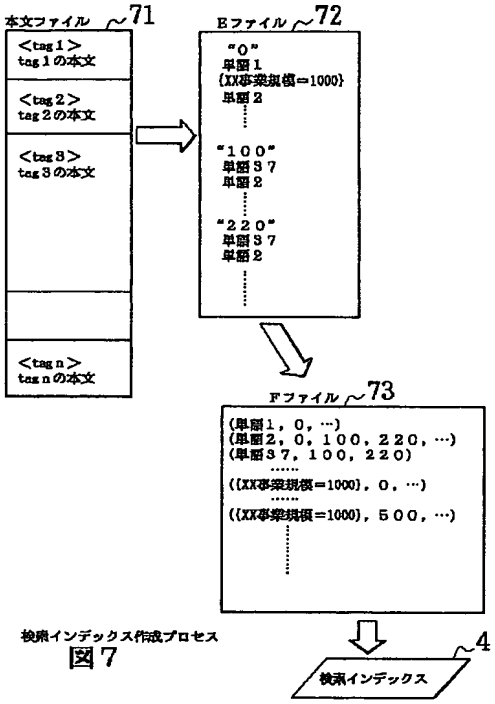
図１１

【図２】



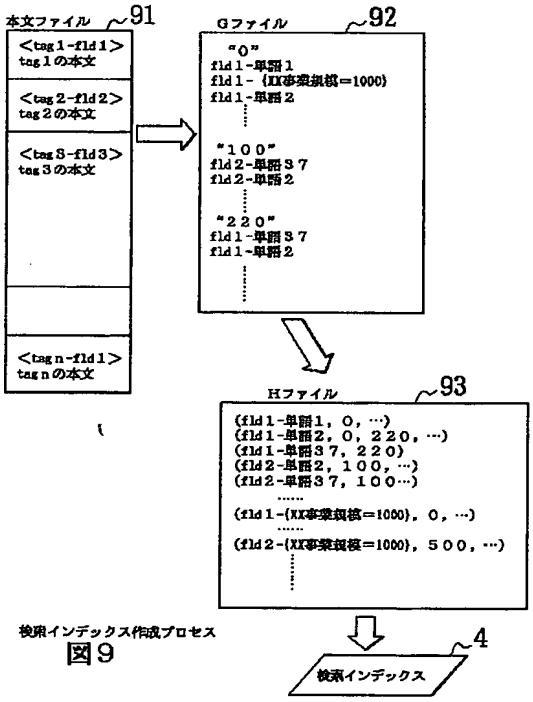
検索インデックス作成処理フロー
図２

【図７】



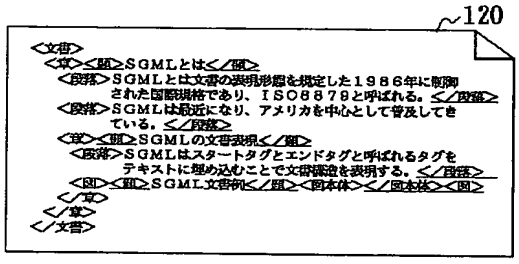
検索インデックス作成プロセス
図７

【図９】



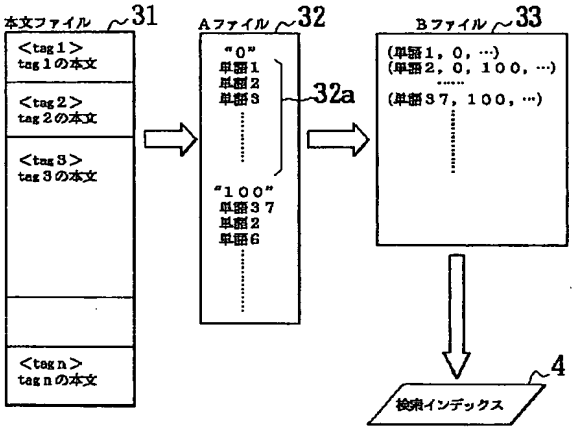
検索インデックス作成プロセス
図９

【図１２】



省略されたタグが復元されたタグ付き文書の例
図１２

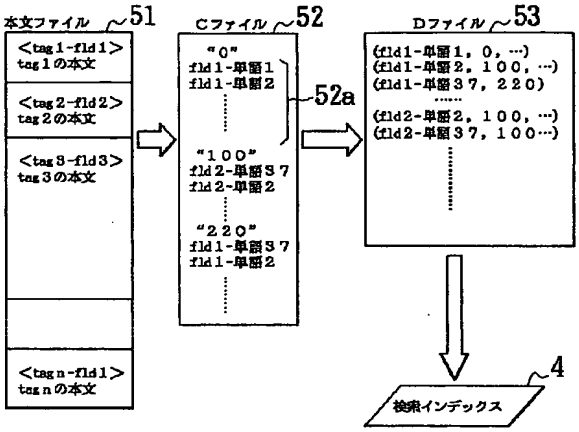
【図３】



検索インデックス作成プロセス

図３

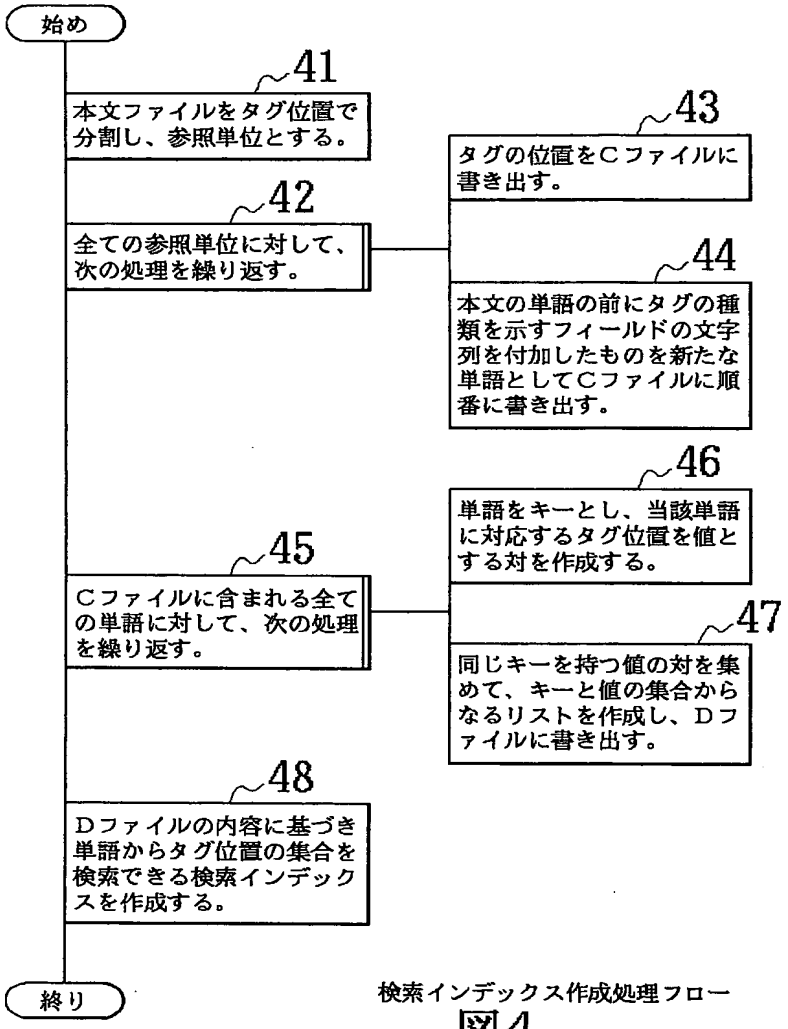
【図５】



検索インデックス作成プロセス

図５

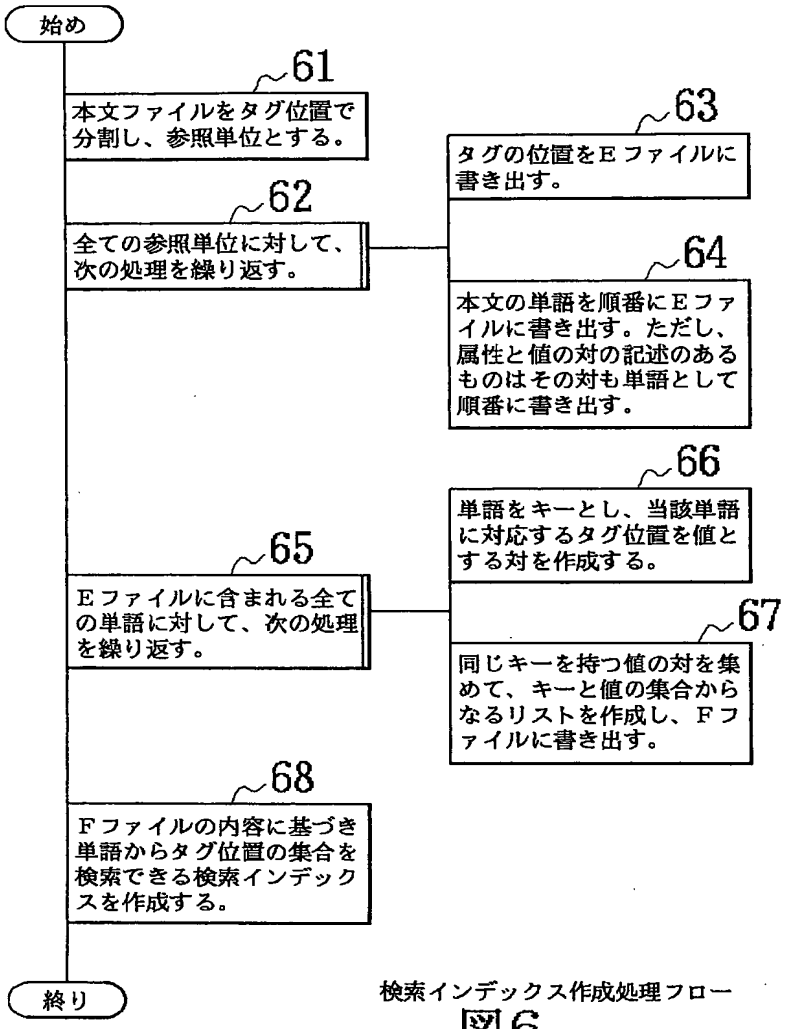
【図４】



検索インデックス作成処理フロー

図４

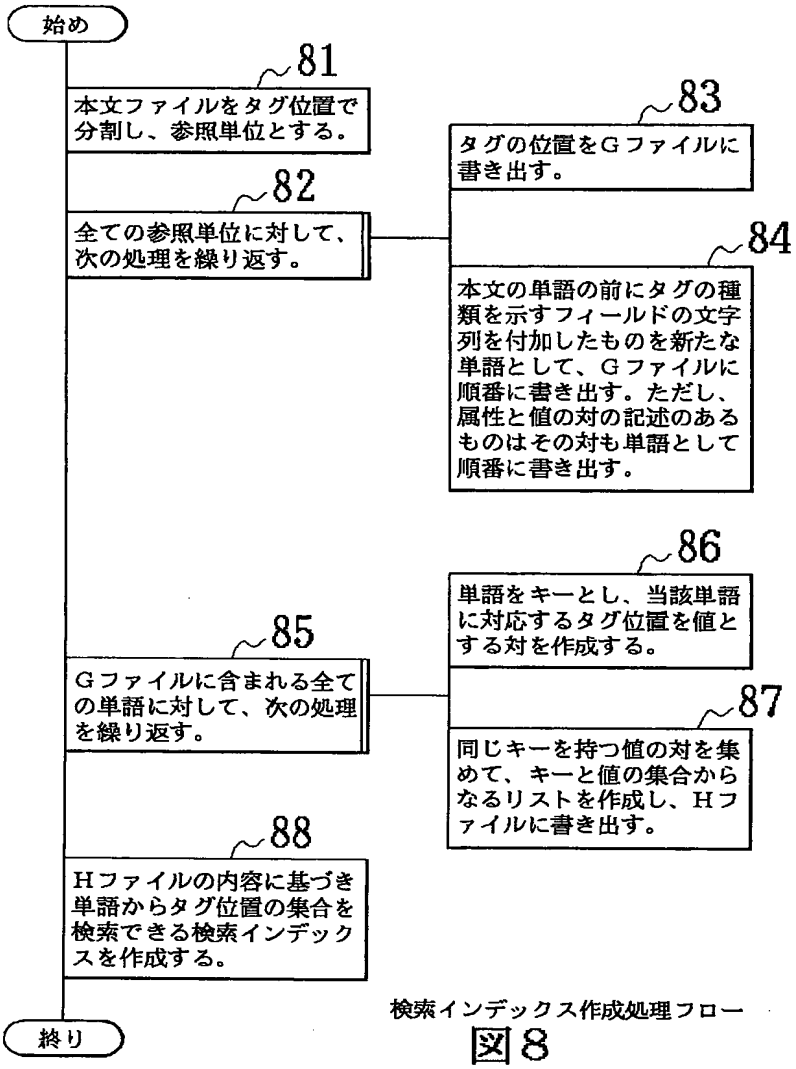
【図６】



検索インデックス作成処理フロー

図６

【図８】



検索インデックス作成処理フロー

図８